DOCUMENT RESUME

ED 164 646 UD 018 890

TITLE Project Outcomes Reporting. Annual Report.
INSTITUTION New York State Education Dept., Albany. Bureau of School and Categorical Programs Evaluation.
PUB DATE Jul 78
NOTE 25p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Data Collection; *Educational Programs; Educational Research; Educational Testing; Elementary Secondary Education; *Program Evaluation; *State Programs
IDENTIFIERS *New York; *Reporting Procedures

ABSTRACT

In this booklet, the New York State Department of Education's approach to the evaluation of categorically aided projects is described. It is stated that in recent years, the emphasis has been on structuring the content of data and limiting it to factors which are comparable and useful at local, State and Federal levels. Measurement procedures which should be employed in the data collection process are detailed. Suggestions are made with regard to language, local testing objectives, norming procedures, functional level testing, evaluation design models, and alternative designs to conventional testing methods. Format and data transmittal systems for aggregating data supplied to the State Department of Education by the schools are explained. A list of guidelines for categorical programs evaluation plans is provided. (WI)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED164646

PROJECT OUTCOMES REPORTING

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

N.Y. State Educ Dept Joan A. Miller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

July 1978

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

UD018890

The University of the State of New York
THE STATE EDUCATION DEPARTMENT
Bureau of School and Categorical Programs Evaluation
Albany, New York 12234

THE UNIVERSITY OF THE STATE OF NEW YORK

Regents of The University (with years when terms expire)

1981	Theodore M. Black, A.B., Litt.D., LL.D., Pd.D., D.C.L., L.H.D. Chancellor	Sands Point
1988	Willard A. Genrich, LL.B., L.H.D., LL.D. Vice Chancellor.	Buffalo
1986	Kenneth B. Clark, A.B., M.S., Ph.D., LL.D., L.H.D., D.Sc.	Hastings on Hudson
1983	Harold E. Newcomb, B.A.	Owego
1982	Emlyn I. Griffith, A.B., J.D.	Rome
1983	Mary Alice Kendall, B.S.	Irondequoit
1984	Jorge L. Batista, B.A., J.D., LL.D.	Bronx
1982	Louis E. Yavner, LL.B.	New York
1979	Laura Bradley Chodos, B.A., M.A.	Clifton Park
1980	Martin C. Barell, B.A., I.A., LL.B.	Kings Point
1981	Joseph R. Bongiorno, B.S., M.D.,	Brooklyn
1984	Louise P. Matteoni, B.A., M.A., Ph.D.	Bayside
1981	J. Edward Meyer, B.A., LL.B.	Armonk
1985	Arlene B. Reed-Delaney, B.A., M.D.	Albany
1987	R. Carlos Carballada, B.S.	Arcade

President of The University and Commissioner of Education
Gordon M. Ambach

Executive Deputy Commissioner of Education
Joseph J. Blaney

Deputy Commissioner for Elementary, Secondary and Continuing Education
Robert R. Spillane

Coordinator of Elementary, Secondary and Continuing Education Opportunity Programs and Urban School Services
Sterling S. Keyes

Director, Division of Federal Education Opportunity Programs
John L. House

Chief, Bureau of Federal Education Opportunity Programs (Upstate)
William C. Flannigan

Chief, Bureau of Federal Education Opportunity Programs (NYC)
Paul M. Hughes

Director, State Education Opportunity Programs
Richard S. Weiner

Chief, Bureau of Migrant Education
Richard A. Bove

Director, Division of Planning Field Services
John J. Murphy

Director, Division of Research
Carl E. Wedekind

Chief, Bureau of School and Categorical Programs Evaluation
Leo D. Doherty

TABLE OF CONTENTS

	PAGE
Introduction	1
Measurement	5
Language	5
Matching Local Objectives To Test Coverage	6
Norming Periods	7
Functional Level Testing	8
Emerging Evaluation Design Models	9
Alternative Designs	11
Format and Data Transmittal	14
Identification	14
Descriptors	15
Results	15
Guidelines For Categorical Programs Evaluation Plans	16

FOREWORD

The Bureau of School and Categorical Programs Evaluation provides services to the Department's program officers for about eight categorically aided programs. The largest of these programs are the State supported Additional Apportionment State Aid (AASA) Program and the federally supported Elementary and Secondary Education Act (ESEA) Title I.

The Bureau reads or reacts to over 4,000 project proposals or plans annually and makes recommendations about the evaluative aspects of projects whose total is in excess of $600,000,000. Through such work, the Bureau deals with each of the State's nearly 750 school districts, many of the Boards of Cooperative Educational Services (BOCES), and other agencies and departments of State government.

In addition to the work described above, the Bureau is engaged in other projects of an evaluative nature, some research and developmental work, a longitudinal study, prediction studies based upon selected school and nonschool factors, and other phases of research and evaluation. However, in regard to categorically aided projects, local district staff deal primarily with George Cronk, William Jaffarian, Eileen Kelly, Robert Miller and George Whitcomb.

Introduction

In recent decades the whole area of educational evaluation has undergone major revision. Ascertaining the relative effectiveness of funded projects seems defensible, especially from the funder's point of view. Some of the new uses for measurement and evaluation in education include: needs assessment, pupil services, readiness to continue, placement, granting of scholarships, diagnosis, devising scholarships, appraising new programs, instructional materials, procedures, programs or others which require relatively hard data for decision-making purposes. On the one hand the field is in its infancy; conversely, since decisions will have to be made, data which may be thought of as developmental in nature are better than no data at all.

The State Education Department's approach to the evaluation of categorically aided projects had evolved from a point in 1966 where local district staff made the majority of decisions about the organization and substance of their annual reports. While content often included both process and product data, the Department soon learned that it could not aggregate the majority of the process data received and, further, that much of the product data were not comparable. As a result the Department began to structure content, concentrate on product data, and limit data to those which were comparable (regardless of district size) and useful both at the local, State and Federal levels. Regarding the latter the major goal of the

Department is to encourage districts to elicit data which will be helpful locally to make the needed decisions. Such data, when aggregated with similar data, will permit the State to make needed decisions about the statewide progress or experience with a given program. Similarly national level staff should be able to aggregate data from all or selected states to assist with decision-making at that level.

After well over a decade of working with variously organized data, the Department reduced its need for product data to the following broad areas:

A- The identification of the learner accompanied by either the appropriate age or birthdate, sex and current grade level.

B- The identification of the treatment by the subject field, the intensity or contact hours of exposure to such treatment, the teaching methodology or activities employed and the location of the treatment (district and building).

C- The "hard data" section indicating whether or not the learner was identified through the use of a standardized screening test, the distribution of the test scores for the whole group and the point below which supplementary education was deemed needed, and the learners' location (in terms of some standard description) on standardized tests, both at the beginning and upon the conclusion of the treatment.

While the three areas of data need may be stated simply there are conditions which tend to make them appear more complicated. In the past some districts elected to complete the resultant computations, thereby increasing the need for resources. Similarly appropriate designs had to be agreed upon, the selection of which sometimes required the accumulation of needed data at specific points in the life of the project. The use of individual data will tend to reduce such burdens. The Department experimented with the collection of individual data in over 200 districts in New York State during the 1976-77 academic year, and in 1977-78 requested that all Pupils with Special Education Needs (PSEN) and Pupils with Handicapping Conditions (PHC) projects report the appropriate codes for individual data. The Department's Bureau of School and Categorical Programs Evaluation works with about eight categorically aided programs, all of which eventually will be requested to report using individual data. ESEA Title I will be requested to report such data for the 1978-79 year.

Many of the designs for assessing project outcomes were originally borrowed from the broad field of research. Since each had its limitations, many were modified to meet decision-making needs. Recently the United States Office of Education developed some models which were felt to be better suited for many of the aspects of ESEA Title I work. However during the pilot test even some of these special models have been found to contain limitations. Economy features in some of the models may be helpful to districts, especially during periods of budget cutbacks. Thus whether or

not the United States Office of Education (U.S.O.E.) will require the employment of only such models, the Department has already afforded districts the opportunity to use them for reporting the 1977-78 year-end data. It is likely that for the school year 1978-79 all districts will be urged to employ one of the newer designs, although the reporting of individual data may reduce the need to be so specific relative to the selection of a model.

Conditions relating to data aggregation should be understood. The Department believes that each district should assess the relative effectiveness of its own projects to enable local administrators to make better decisions. However the Department's main activity is one of aggregating data for similar projects rather than assessing the effectiveness of each local project. It can be seen that, with individual data, either approach may be employed, but the Department's main effort will be to try to reduce ineffective projects and aggregate more meaningful state results. For the latter projects with small numbers may often be classified as indeterminate locally, but could become worthy of dissemination when aggregated with other, similar project results statewide.

Measurement

The acquisition of comparable data suitable for aggregation requires that the measures employed be administered under similar conditions across the target population. While slight variations may prevail, the divergence from uniformly agreed-upon test administration and analysis conditions would introduce error that would render the district evaluation data of lesser value at all levels of interest (local, State, and Federal). Local education agencies must be concerned with the language of scores, the boundaries of the norm-referenced groups used for comparisons, the time in the year of the norm-referencing, the functional skill level of the compensatory education pupil, the analytic procedure to which the data will be subjected and the design alternatives when the conditions for the preferred methods simply cannot be met.

Language

Norm-referenced scores are interpreted for different purposes with regard to several aspects of the normal curve. Traditionally compensatory education students were selected for an instructional supplement because they placed below the 23rd percentile rank with reference to their age/grade norm on a particular skill. The

Federal Government devised a metric called the Normal Curve Equivalent (N.C.E.) which was similar to percentile ranks, but had the advantage of being suitable for averaging. The N.C.E. can be easily converted to a percentile rank for discussion with teachers, support staff, parent advisory councils, news media personnel, school boards, and parents. Most test-scoring services provide pupil scores in percentile ranks and normal curve equivalent units. Districts which contract for commercial scoring should request that the normal curve equivalent score be reported back to the district. If the N.C.E. is not available, then the requested score should be the percentile rank. A conversion table to change percentile ranks into Normal Curve Equivalent units is available from the State Education Department. The language of the scores in the late 1970's and the 1980's will be the normal curve equivalent unit both with respect to averages at a point in time and changes in performance.

Matching Local Objectives To Test Coverage

Commercial test publishers generally base the selection of the content of their tests upon the frequency of the content's appearance in instructional materials and the associated objectives. Traditionally publishers surveyed content in the several major basal readers, made frequency tabulations, summarized by objective topic the most frequently selected topics, wrote items, and commenced the standardization procedures followed by a norming process. While the publishers' procedures are defensible, the need to select a nationally normed test with content sampling that matches the district's curriculum and objectives must be recognized. Since most norm-referenced tests provide an item analysis table (items correspond to

objective or skill topics), school districts should match by level the district's objectives with the publisher's objectives list. If the skill content coverage match between publisher and district is not high (70 percent or better), the district is not assessing its own objectives well and low scores might reasonably be anticipated. The effectiveness (or absence thereof) of a compensatory education treatment might not be detected if a widely respected norm-referenced achievement test, which did not sample the skill content in the district well, were administered.

Norming Periods

Test publishers administered the tests to clustered, stratified, randomly selected groups of children at particular times in the year. Most of the widely administered norm-referenced tests were normed in October and May. School districts should administer tests to the target population on the norming dates upon which publishers established the norms. Testing before a norming date tends to depress scores while testing after the norming date inflates the scores. Testing at times other than the norming date, when the norm group is considered to be the comparison group, introduces unnecessary error that could make the interpretation of the collected data suspect. Evidence indicates that extrapolated and interpolated norms are not satisfactory for interpretation. Consequently, the purchaser of nationally standardized tests should be sure that the norming population was similar to the pupils in the school district before a test is selected, but should be sure of the presence of an empirically derived norm from actual student scores at an appropriate point in the overall evaluation design.

Functional Level Testing

The target population in compensatory education treatments tends to fall at the lower end of the normal distribution at a particular grade. Achievement scores tend to be highly reliable for the average-scoring pupil, but become less reliable for participants who score at some distance from the norm. In fact when some compensatory education pupils take tests designed for their age/grade level (so-called "on grade level") the pupils obtain the lowest score possible. This "floor effect" means that the true score of such participants is not known. The experience is also unpleasant for the participant. To remedy the problem of not obtaining reliable information with an on-grade level test, such pupils are sometimes administered out-of-level tests, designed for a grade level below that of the participant. In this way a functional level score may be obtained that reflects the skills acquired by the learner at that point in time. The difficulty, however, is that the out-of-level score must be converted to an on-level test score for purposes of interpretation. The school district must be certain that conversion tables are available from a publisher before a test is purchased for out-of-level administration (functional level testing). Functional level testing may be desirable for some participants in some evaluation designs, but the test purchaser must be careful that either appropriate conversion tables are obtained or the scoring service will transform the scores for on-grade level interpretation.

Emerging Evaluation Design Models

Two models (with variations) are emerging as the principal designs that will yield data suitable for aggregation. Both models require an estimate that defines the performance expected of compensatory education pupils if no compensatory treatments had been implemented.

Model A-1, the norm-referenced model, requires that all compensatory pupils be selected on a measure, usually a test score obtained the previous spring, pretested on the publisher's norming date (usually an October administration), and posttested on the publisher's noming data (usually a May administration). Frequently the spring posttest is a district wide test administration which also furnishes the selection measure for determining eligible pupils for the subsequent year. The Model A-1 design requires that the same form and level of the test be administered at both the pretest data collection point and the test data collection point. Without treatment the target pupils are expected to score at the same percentile ranks (as converted to normal curve equivalents) at the posttest as they scored at the pretest. Differences between expected and obtained distributions are subject to analyses for inferential purposes so that where true differences occurred they may be recognized. The analysis will be carried out by the State, but it is critical that the school district adhere to the data collection schedules and rules stated.

Model C-1, the special regression model, requires that compensatory education and noncompensatory education pupils be tested at the same times with the same form and level of a test, preferrably on the publisher's norming dates (in the spring). All pupils are ranked in descending order by normal curve equivalent unit scores. The compensatory education pupils are selected strictly in order of their deficiency, with the poorest-performing pupil receiving first selection, next poorest performance pupil being the second pupil selected and so forth until the spaces reserved for compensatory education pupils are filled. The score obtained by the last compensatory education pupil selected is denoted as the cut-off score and all pupils scoring above the score are described as comparison group pupils. The needs assessment selection test score and pretest score are both derived from this single test administration. A second test administration (on the publisher's norming date the following spring) provides the second score for the regression line derived from correlating the two scores for each pupil. The no-treatment-expectation is obtained from the comparison group's posttest-on-pretest regression line. A difference between the treatment group and the comparison group is detected where the two regression lines intersect the cut-off score and again at the treatment group's mean pretest score. The difference between expected performance and obtained performance constitutes the normal curve equivalent gain. The State Education Department will provide the statistical analysis for Model C-1 so local education agencies need only conform to the data collection requirements and reporting requirements. Model C-1 may be particularly attractive since achievement tests might be administered once a year in the appropriate subject matter area, thereby saving both instructional time and money. The limitation of Model C-1 is that one

particular level of a test in a district-wide grade or grade in a building that best distributes the whole population (compensatory education plus comparison groups) must be administered. This requirement may rule out functional-level testing for the majority of the compensatory education pupils. Additionally there must be at least 50 comparison group pupils available for each grade's treatment group.

Alternative Designs

Some populations may not meet the requirements necessary to be considered similar enough to the norm groups of most norm-referenced achievement tests. Some compensatory education treatments serve learners who are not available at test administration time. Such occasions have occurred with bilingual pupils, pregnant teenage girls, drug abuse pupils, neglected children, prisoners, delinquent pupils, some handicapped pupils, and children of migrant workers. On other occasions treatments are designed and installed well after a publisher's fall norm date has passed. These circumstances may prohibit the use of the U.S.O.E. evaluation design models.

The most attractive approach to meeting the need to evaluate changes in behavior during compensatory treatment when norm referenced evaluation designs may not be appropriate is to (1) define the behavioral objectives involved in the proposed treatment, (2) match the district-developed objectives with a commercial publisher's bank of objectives, and (3) administer only appropriate objective-

referenced or criterion-referenced test items from the commercial publisher's standardized criterion-referenced test. It is imperative that the district purchase only standardized objective referenced tests since the items that support an objective constitute the test, unlike norm-referenced tests where the whole subtest in a domain (e. g., comprehension) constitutes the test. Standardization with objective-referenced or criterion-referenced tests is specific to the entities of reliability and validity. Such aspects of standardization include the (1) objective to test item link rationale, (2) number of test/items per objective, (3) mastery assurance methodology, (4) degree of difficulty, (5) item equivalence, (6) temporal stability, (7) instructional dependence, (8) validity, (9) hierarchical arrangement and (10) population applicability.

Generally the data collection schedule will involve administering criterion-referenced tests at the outset of a treatment and portions of the criterion-referenced tests as posttests as the pupil indicates a readiness to demonstrate mastery of the objective. The accumulation of objectives by instructional level would be necessary for a summary evaluation report at the end of the treatment. Group analyses tend to be descriptive and presented in tabular form with reference to mastery and non-mastery status.

Measurement activities will continue to emphasize behavioral objectives either in the form of skill mastery for objective/criterion-referenced measures or gains in normal curve equivalent units for norm-referenced tests. School districts should pay close attention to matching the objectives undergirding the measures with those embraced by the district. Functional level testing will impart more information about the educationally disadvantaged learner when USOE Model A-1 is selected as a design. Considerable savings in time and money are possible with the selection of USOE Model C-1 evaluation design. The State Education Department will conduct an analysis suitable for the aggregation of comparable data when (1) data are provided by individual student in the format described in the next section and (2) the data are collected according to the requirements of the evaluation models. Where data cannot be collected according to the requirements of the evaluation models, local education agencies should develop evaluation designs that collect achievement data through the use of standardized criterion-referenced tests.

Format and Data Transmittal

To facilitate the aggregation of data supplied by school districts to the State Education Department, such data must be reported uniformly. Evaluation report forms (incorporating the format for key punching and magnetic tape reporting) with detailed instructions are supplied to each district. Information elicited from the districts falls into three areas, as follows.

Identification

In each table of the report form, the district code, Board of Education (BE) function (NYC only), and Basic Educational Data System (BEDS) school building code are entered. Where individual data only are being reported, as is currently the case with Pupils with Special Education Needs (PSEN) and Pupils with Handicapping Conditions (PHC) reporting (and as will be eventually for all programs) a permanent pupil number is also entered. This number, assigned to the pupil by the district, may not exceed eight alpha/numeric characters and is permanent. The pupil retains this same number for all components or other compensatory programs in which the pupil receives treatment until he or she is graduated. Pupils are never identified by name and the key to their identity is retained by the district to ensure confidentiality. The date of birth and current grade level are also reported, as well as identification of the pupil as educationally disadvantaged or in a category of handicapping conditions.

Descriptors

Further entries in the tables indicate the subject field, teaching methodology or activity, objective (desired outcome), test instrument administered, number of pupils treated and tested, number of hours of exposure to treatment, etc. All such information is reported in numeric code form. These codes are included in the instructions sent to the districts.

Results

In the same format, the district reports the pupils' test scores or objective-based assessment at the beginning of treatment and at its conclusion. Where the information is reported on an individual basis, the raw score and percentile rank are submitted. Submission of group data by the district necessitates reporting some additional data, such as the test used to determine the level of statistical significance and the obtained value of this test statistic.

When the district submits individual data it no longer performs any computations or analysis of data to fulfill reporting requirements. While most districts are encouraged to and actually do analyze their data to determine the effectiveness of their programs it no longer is a requirement. The Education Department has elicited enough individual information to aggregate data for all districts and enable it to assess the statewide effectiveness of the various programs.

Guidelines For Categorical Programs Evaluation Plans

1. Project objectives must be stated in behavioral or measurable terms.
2. Objectives must describe how the learner's behavior will be changed by the treatment (activities).
3. Objectives must indicate how much change constitutes improvement.
4. Performance criteria to verify that change has taken place must be stated.
5. Where pupils receive other similar treatments, a comprehensive evaluation design across all such programs must be included.
6. Data collection procedures should be specified.
7. All treated participants must be tested.
8. Each participant must be tested at least once annually, or where necessary, twice annually.
9. Plans for reporting individual data must be included.

10. Only standardized instruments (e. g., norm-referenced tests, criterion-referenced tests) may be used for evaluative purposes. Standardized instruments are those devices which have objectively based evidence indicating that the instrument measures the trait(s) it purports to measure and that the results from such measurement are relatively consistent over a period of time when there has not been a change in the trait.

11. Generally norm-referenced tests with norms developed more than 8 years ago are not acceptable.

12. When using a norm-referenced test, the level of the test administered to a pupil should correspond to the pupil's level of performance and not necessarily to the pupil's grade or age level. (Out-of-level testing.)

13. Where the participants receive instruction for the school year and when testing takes place only once a year, the test should not be administered prior to May nor later than the end of June.

14. Where the participants receive instruction for the school year and when testing takes place twice a year, the pretest should not be administered prior to September or later than the end of October. For the posttest item 13 above applies.

15. A norm-referenced test must be administered in the month(s) in which it is normed. Tests which provide extrapolated or interpolated norms as they relate to items 13 and 14 above are unacceptable for use in program evaluation.

16. Dates of pretest and posttest administration must be specified.

17. For short duration programs, particularly those which provide less than 60 hours of instruction per pupil, a criterion-referenced test must be used.

18. Generally criterion-referenced tests should be used to evaluate PHC programs.

19. A proposal to use unpublished instruments must be accompanied by statistical evidence including instrument reliability and validity indices.

20. The form(s), level(s), and publication date(s) for standardized instruments must be specified.

21. Subtests for reading and/or mathematics must be specified.

22. The pretests and posttests must be compatible test instruments.

23. When a readiness instrument is used as a pretest an achievement instrument can not be used as a posttest. Rather a readiness instrument must be used as both a pretest and posttest measure. The same principle applies to achievement test usage.

24. Both raw scores and percentiles are needed for reporting norm-referenced tests.

25. Acceptable plans for the evaluation of objectives must be included.

26. A plan for data analysis should be provided.